Opinion

Is mass spectrometry ready for proteome-wide protein expression analysis?

Juri Rappsilber and Matthias Mann

Address: Protein Interaction Laboratory in the Center of Experimental Bioinformatics, Department of Biochemistry and Molecular Biology, University of Southern Denmark, Campusvej 55, DK-5230 Odense M, Denmark.

Correspondence: Matthias Mann. E-mail: mann@bmb.sdu.dk

Published: 31 July 2002

Genome ***Biology*** 2002, **3(8)**:comment2008.1–2008.5

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2002/3/8/comment/2008

Abstract

Recent advances in mass spectrometry will soon allow routine analysis of protein expression levels. How close are we to true quantitative proteomics?

Although the information content of the genome is static, it is changes in the molecular composition of cells that govern life. The abundance of many transcripts undergoes extreme fluctuations, for example during the cell cycle or in cellular responses to external events. Studies that use microarrays or PCR-based approaches aim to quantify all the transcripts in a cell in order to account for its phenotype and to provide a basis for understanding cellular processes. But the assumption that up- and down-regulation of mRNA accompanies functional changes in the cell does not hold in all instances. Some studies focus on the analysis of polysomal mRNAs, so as to examine the translationally active fraction of the mRNA pool; although this brings the analysis of mRNA closer to cellular function, it is proteins, not mRNAs, that effect most processes in the cell.

Regulatory mechanisms exclusively affecting proteins include the inhibition of translation by the binding of regulatory proteins to mRNA, translational hopping, co- and post-translational modifications, proteolytic processing, co-factor binding, and the localization of proteins within the cell. Knowledge of the proteome of a cell will therefore allow investigators to obtain a more representative picture of the cell at the molecular level. The proteome is defined as the time- and cell-specific protein complement of the genome - so it encompasses all proteins that are expressed in a cell at one time, including isoforms and post-translational modifications. In this article, we conclude that proteomic analysis by mass spectrometry will soon be capable of competing with or complementing the use of mRNA chips for the analysis of expression levels.

Proteomics

The complexity of proteomes precludes addressing all their parameters in a single experiment. Several approaches are designed to address particular aspects of proteomes, such as the expression level of proteins - their presence and abundance - or the presence of sequence isoforms in a given cell type, under certain experimental conditions, or among different cell states [1-3]. Alternatively, cellular compartments can be isolated, to build an inventory of their protein complement, or protein-protein interaction maps can be built from the systematic purification of multi-protein complexes and the identification of their members [4-6]. Moving beyond basic protein identification, current investigations also aim to describe all protein modifications and their changes in response to various challenges to cells.

The idea of proteomics dates back to an initiative to systematically archive the patterns of protein spots seen on two-dimensional electrophoresis gels with samples from different cell types, fluids, and organisms [7]. More recently, the development of mass spectrometry has brought the sensitivity and speed required to obtain sequence and modification data about the proteins contained in two-dimensional gel spots. Mass spectrometrists investigate the peptides obtained by cleavage of the protein within a spot using proteases specific to certain amino acids within a protein sequence, such as trypsin. The measured masses of the cleaved peptides are compared with theoretical digests of the known and predicted proteins contained in databases, in a process of ‘peptide mass fingerprinting’. Furthermore, any observed peptide species can be selected and fragmented further within a tandem mass spectrometer to yield precise sequence information. In this step, characteristic fragments and mass differences also reveal protein modifications. One caveat of coupling two-dimensional gel electrophoresis to mass spectrometry is, however, that it results in a biased measurement of the proteome, because there are inherent limitations of two-dimensional gels in displaying proteins at extremes of isoelectric point (pI), size, or hydrophobicity. Also limiting is the dynamic range of two-dimensional gels, which is the ratio of the most abundant to the least abundant protein visible in one experiment. Even though several thousand spots can be seen on a good two-dimensional gel, multiple isoforms and modification stages derived from a single gene may account for many of them. Thus, even the best two-dimensional gels usually represent the products of only a few hundred genes (see, for example, [8]). It should also be noted that this approach is very labor-intensive and, although it is amenable to automation, it cannot compete in terms of throughput with other functional genomic techniques such as microarrays or two-hybrid protein interaction mapping.

Tackling complex protein mixtures

It was shown as early as 1992 that thousands of peptides could be analyzed by liquid chromatography coupled to tandem mass spectrometry (LC MS/MS) [9]. Continuous advances in the methodology and automation of both LC MS/MS and database searching now allow the identification of large numbers of gene products [10]. To achieve this, a protease is usually added to the protein solution to produce peptides from all the proteins in the sample at once, rather than first separating the individual proteins on a gel and digesting them individually. A major advantage of this procedure over two-dimensional gel electrophoresis is that the detection step is automated and is achieved in the same step as the identification of the protein. Furthermore, all proteins are represented, rather than only those visible on gels. Another advantage is that the analysis is less biased against certain classes of proteins (large, small, acidic, basic or hydrophobic). The complex mixture of peptides is separated first in a liquid chromatography step then further in the on-line coupled tandem mass spectrometer to yield isolated peptides. The fragment patterns obtained from the peptides allow identification of the genes from which the components of the original protein mixture are derived. Common peptides from all the individual products of a particular gene are detected together and thus increase the chance of detecting products of that gene. The trade-off is, however, that no estimate of the number or nature of variants can be obtained, as would come from the protein masses taken from two-dimensional or one-dimensional gels. Also, a comprehensive characterization of a complete protein remains the exception.

Using two-dimensional liquid chromatography to separate the peptides - with the two dimensions reflecting two independent peptide properties, such as number of charged amino acids and hydrophobicity - prior to mass spectrometry allows the analysis of even more complex samples. Such analyses are proving to be a substantial improvement over two-dimensional gel analysis. A recent investigation of a total lysate of the yeast *Saccharomyces cerevisiae* using LC MS/MS claims the identification of nearly 1,500 proteins [10]. In contrast, one of the most successful two-dimensional gel analyses to date was done with the bacterium *Haemophilus influenzae* and revealed the presence of protein products from only 502 genes [11].

Stable isotopes and chemical modifications

The value of recent advances in chromatography and mass spectrometry has been built upon by the introduction of stable isotopes into two samples - for example by growing cells on normal and stable-isotope-enriched media - to allow a quantitative analysis of the differences between the samples. For example, in a new method termed SILAC (for isotope labeling by amino acids in cell culture) one amino acid is used either in the normal or in a stable-isotope-containing form, to distinguish the proteins of states A and B [12]. The combined samples are fractionated, digested, and analyzed by LC MS/MS. As isotopes differ in their mass, a peptide from one state will be a defined number of mass units different from the same peptide from the other state, resulting in a signal doublet in the mass spectrum. The relative signal intensity within the doublet indicates the ratio of protein concentrations in the two samples. Chemical selection of peptides containing a specific amino acid, for example cysteine, reduces the complexity of a peptide mixture and so should permit the analysis of somewhat more complex samples. It should be borne in mind, however, that only a subset of proteins contains the specific amino acid in a peptide that falls into the mass range accessible to mass spectrometric analysis.

Introducing a chemical group for specific purification allows a label to be introduced for the relative quantitation of peptides from different samples; a heavy reagent modifies one sample, a light reagent the other (typically by replacing a number of hydrogen atoms by deuterium or ^{12}C by ^{13}C). As before, a mass shift allows simultaneous detection and quantitation of identical peptides from different samples. Different peptides are detected with greatly different signals because of differences in the loss due to adsorption to surfaces and in ionization efficiency, allowing detection in the mass spectrometer. If the only difference is in an isotope that affects only a few atoms, however, the ratio of the peak areas is an accurate measure of the relative abundances. Our experience indicates that quantitation to better than 20% in both accuracy and precision can be achieved - that is, detection of a 20% change in the expression level of a single gene product (see Figure 1 for an example of a typical experiment). This concept is now known as stable isotope quantitation, and if it incorporates an affinity tag it is known as isotope-coded affinity tags (ICAT) [13].

One of the first applications of the ICAT method was the quantitative profiling of differentiation-induced microsome-associated proteins in myeloid leukemia (HL-60) cells [14]. HL-60 cells differentiate upon treatment with 12-phorbol 13-myristate acetate (PMA). Lysates of control and PMA-treated cells were fractionated by differential ultra-centrifugation and microsomal fractions were isolated; as the microsomes are derived from the endomembrane systems of the cell, this isolation step is expected to enrich for membrane-bound or secreted proteins. After denaturation with the detergent SDS, isotopically normal (control) and heavy (PMA-treated) forms of the sulfhydryl-specific ICAT reagent were used to label all the cysteines in the protein population. After the labeling reaction the samples were combined, digested with trypsin and subjected to a three-step chromatographic purification and separation of the affinity-labeled peptides. First, cation-exchange chromatography cleaned the peptide mixture of detergent and of excess reagents, and this resulted in 30 eluted peptide fractions. Avidin columns were then used to affinity-purify labeled peptides from each fraction. Finally, the recovered, labeled peptides were separated and analyzed by LC MS/MS. A suite of software tools attempted to streamline the analysis and interpretation of the recorded mass spectra. The procedure resulted in a total of 491 identified and quantified proteins, of which only 50 were membrane proteins (despite the starting material having come from a microsomal fraction) [14].

It may still be premature to expect large amounts of data from this technique, as a number of critical steps await optimization. The protocol for affinity purification of the labeled peptides requires improved robustness in selectivity as well as in binding and in elution efficiency. On the analysis side, the main limitation of the mass spectrometry lies in its automation, as software for data-dependent analysis and quantitation is only now emerging. Overcoming these limitations is, however, within reach.

LC MS/MS - what is technically feasible?

The sensitivity of LC MS/MS has improved dramatically within the last five years: it can now identify 1-10 fmol of a peptide loaded on a chromatography column in routine analysis. Analysis of a protein that is present at a level of 1000 copies per cell would theoretically require 10^6 cells to start with, assuming no losses during protein purification. Soon, improved instrumentation might allow detection of 10 amol - this is 6,000,000 particles, and means that 10^5-10^6

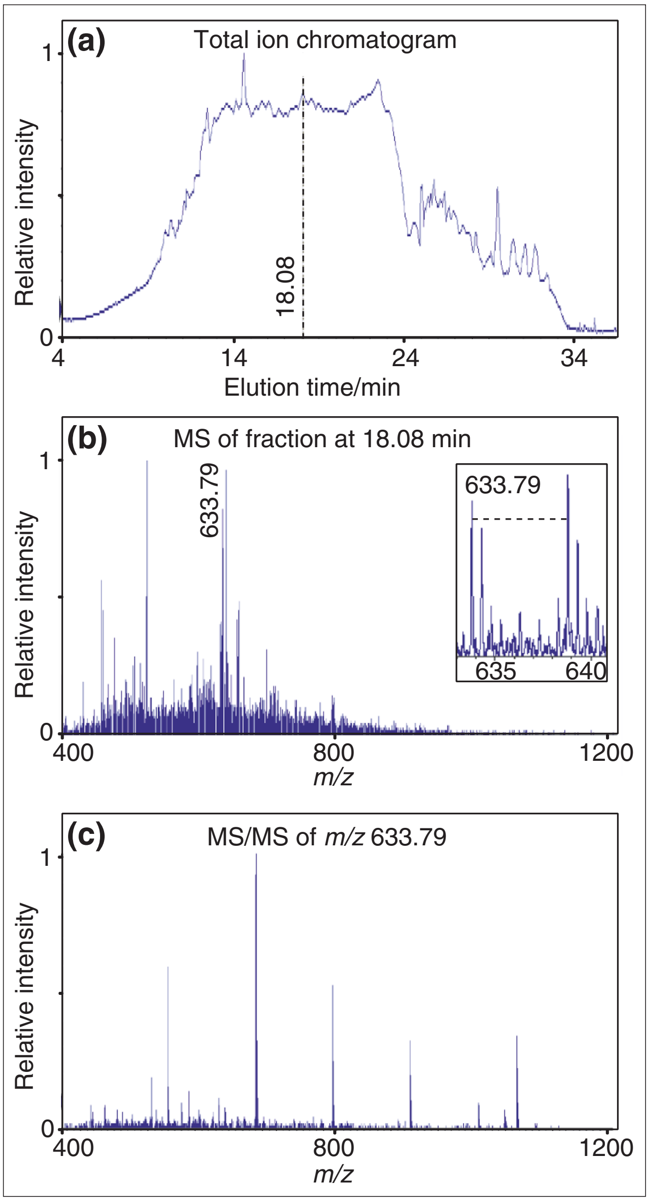

Figure 1
LC MS/MS analysis of a peptide mixture obtained by proteolysis of a complex protein mixture. **(a)** The total ion chromatogram (sum of the signals of all peptides as a function of time) of peptides obtained by tryptic digestion of a complex protein mixture is shown, as obtained by liquid chromatographic separation and on-line analysis by mass spectrometry. **(b)** As an example, the mass spectrum of peptides eluting at 18.08 minutes after the start of the analysis is shown. **(c)** The doublet starting at 633.79 indicates a pair of peptides that are chemically identical yet differ in mass due to stable isotope incorporation in the two different samples. Their relative intensities reveals the relative abundance of the protein from which they derive in the two protein populations that were mixed - in this case approximately 0.9:1.0. The protein is then identified by fragmentation of the peptides. *m/z* is this the mass/charge ratio.

cells would be sufficient to analyze proteins that are present at levels of only ten copies per cell [15].

Protein concentrations in cells vary over approximately seven orders of magnitude and in blood or serum can even vary over a range of 10^{12}. LC MS/MS itself has a dynamic range of four orders of magnitude; by protein fractionation this can be increased about one hundred-fold to 10^6. The technique in principle, therefore, allows a near-exhaustive analysis of most protein samples. And the time required and the amount of data generated in the analysis of a total cell lysate are not insurmountable. Currently, it takes about one second to sequence a peptide. In order to increase confidence two peptides should be sequenced for each protein. If one assumes 10,000 genes to be active in a human cell, it would under ideal conditions take 20,000 seconds or around 6 hours to analyze their products. Of course, digestion will result in more than two peptides per protein; some proteins will contribute one hundred or more peptides to the sample, and sequence diversity including modifications will further increase the number of distinct peptides. To reduce sample complexity, approaches using amino-acid-specific reagents like ICAT depend on selective purification of only a subset of peptides from each protein. The remaining complexity can be addressed by intelligent software that prevents the sequencing of peptides that come from proteins that have already been identified. It is even possible that the time-consuming sequencing of peptides can be omitted for the majority of peptides: a database containing elution times and peptide masses could be used to identify eluting peptides on the basis of accurately determined mass and elution time alone. In this case, even more peptides could be identified and quantified, theoretically up to as many as dozens per second.

Clearly, analyses of complex protein mixtures on the basis of only a few peptides per protein gives limited access to the diversity of protein products that can result from a single gene. The more proteins are present in a sample, the less information can be obtained about each one. In highly complex mixtures, therefore, one strives to identify as many proteins as possible and to catalog a subset of the modifications present. In contrast, if highly enriched proteins are available, for example if a specific antibody purifies the protein products of a single gene, much more extensive protein characterization is possible, including near-exhaustive detection of splice variants and protein modifications.

Proteins or mRNA?

Once tools for the quantitative analysis of mRNAs and proteins are implemented, the inevitable question arises as to which of them is better suited for a given application. Microarrays, as well as PCR-based methods, have the potential to be extremely sensitive. Currently 1-10 μg of total RNA, corresponding to 10^5-10^6 cells, are required for global expression analysis by DNA arrays, but in future a few hundred cells might be sufficient, and if only a small number of genes is investigated, this detection limit can already be achieved today. It is possible to limit an experiment to those genes that are of special interest, by designing an appropriate DNA chip or specific primers for amplification, providing a benefit for directed studies. As the procedures for such experiments are easily automated, these are likely to be the first methods to be attempted in most cases. Microarray probing can be done in a highly parallel manner, making it a suitable tool for screening, especially as costs decrease.

Protein expression analysis by mass spectrometry is closing in on cDNA-based approaches in terms of sensitivity and accessibility. Although PCR allows amplification and therefore highly sensitive analysis of specifically targeted mRNAs, a global analysis requires similar amounts of cells for quantitation of mRNAs or proteins. The strength of mass-spectrometric quantification during protein-expression analysis is accuracy and precision. As mentioned above, changes of 20% in expression level can be detected with high reproducibility by mass spectrometry, whereas chip-based methods show a production-introduced variability that demands repeated experiments and usually allows the recognition of only two-fold or greater changes in mRNA abundance. It should be noted, however, that special instrumentation, in the form of mass-spectrometric equipment, as well as the required chromatographic tools, will require higher specialization by the laboratory user than chip-based methods.

None of the methods targeted at mRNA will, however, be able to monitor events that occur post-transcriptionally. The investigation of proteins allows the study of post-transcriptional regulatory mechanisms and their effects. Events considered to be rare, such as translation hopping, may turn out to be more frequent once their detection is easier. Better-known regulatory mechanisms include the post-translational modification of proteins that mediate key functions, such as signal transduction, and the ubiquitinylation system (attachment of ubiquitin) that targets proteins for destruction. With proteomic methods, in contrast to mRNA-based methods, it is also possible to follow changes in protein localization. The nuclear matrix, cytosol, plasma membrane, and other cellular fractions can be isolated and analyzed separately to follow changes in their protein composition. This affects, for example, proteins that are stored in the cytosol to be shipped into the nucleus for action: subcellular localization changes while the total amount of the protein remains constant. By similar logic, it would be very desirable to study temporal changes in protein-protein interactions that result, for example, in changes in the composition of dynamic multi-protein complexes, such as the spliceosomal complexes during splicing of pre-mRNA or the centrosome during the cell cycle. Proteome analysis will be the method of choice in all these cases. It is already the only possibility for the analysis of mRNA-free samples, including cell-free body fluids; it is worth noting the ease of sampling body fluids, making them ideally suited for diagnostics.

A general problem of large-scale analyses is that they can generate an overabundance of data. After data acquisition, false positives must be sorted out in a resource- and time-consuming follow-up study. For this reason, it is beneficial to start with techniques that result in the most relevant data and the shortest list of false positives. We would argue that such techniques would focus on proteins, the end points of gene expression and of all regulatory events. Furthermore, proteins and not mRNAs are the targets of most drugs, adding value to the results of investigations that directly address proteins. Mass spectrometry has already proven to be an invaluable tool for the identification of proteins and their modifications, and it will soon be capable of quantitative expression analysis in many instances. We are close to being able to fulfill the promise of proteomic techniques and to having a complete picture of the types, and abundance levels of gene products in particular cell types, developmental stages and locations within the body; the future for proteome-centered biology and medicine looks bright indeed.

Acknowledgements

We thank members of our group, especially Leonard Foster, Carmen de Hoog and Shao-En Ong, for critical reading of this manuscript. J.R. is a Marie Curie Fellow. The work in our laboratory is supported by generous funding from the Danish National Research Foundation to the Center of Experimental Bioinformatics.

References

1. Griffin TJ, Aebersold R: **Advances in proteome analysis by mass spectrometry.** *J Biol Chem* 2001, **276:**45497-45500.
2. Mann M, Hendrickson RC, Pandey A: **Analysis of proteins and proteomes by mass spectrometry.** *Annu Rev Biochem* 2001, **70:**437-473.
3. Rappsilber J, Mann M: **What does it mean to identify a protein in proteomics?** *Trends Biochem Sci* 2002, **27:**74-78.
4. Neubauer G, King A, Rappsilber J, Calvio C, Watson M, Ajuh P, Sleeman J, Lamond AI, Mann M: **Mass spectrometry and EST-database searching allows characterization of the multi-protein spliceosome complex.** *Nat Genet* 1998, **20:**46-50.
5. Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM, *et al.*: **Functional organization of the yeast proteome by systematic analysis of protein complexes.** *Nature* 2002, **415:**141-147.
6. Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutilier K, *et al.*: **Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry.** *Nature* 2002, **415:**180-183.
7. Appel RD, Sanchez JC, Bairoch A, Golaz O, Miu M, Vargas JR, Hochstrasser DF: **SWISS-2DPAGE: a database of two-dimensional gel electrophoresis images.** *Electrophoresis* 1993, **14:**1232-1238.
8. Fountoulakis M, Juranville JF, Berndt P, Langen H, Suter L: **Two-dimensional database of mouse liver proteins. An update.** *Electrophoresis* 2001, **22:**1747-1763.
9. Hunt DF, Henderson RA, Shabanowitz J, Sakaguchi K, Michel H, Sevilir N, Cox AL, Appella E, Engelhard VH: **Characterization of peptides bound to the class I MHC molecule HLA-A2.1 by mass spectrometry.** *Science* 1992, **255:**1261-1263.
10. Washburn MP, Wolters D, Yates JR: **Large-scale analysis of the yeast proteome by multidimensional protein identification technology.** *Nat Biotechnol* 2001, **19:**242-247.
11. Langen H, Takacs B, Evers S, Berndt P, Lahm HW, Wipf B, Gray C, Fountoulakis M: **Two-dimensional map of the proteome of *Haemophilus influenzae*.** *Electrophoresis* 2000, **21:**411-429.
12. Ong SE, Blagoev B, Kratchmarova I, Kristensen DB, Steen H, Pandey A, Mann M: **Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics.** *Mol Cell Proteomics* 2002, **1:**376-386.
13. Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH, Aebersold R: **Quantitative analysis of complex protein mixtures using isotope-coded affinity tags.** *Nat Biotechnol* 1999, **17:**994-999.
14. Han DK, Eng J, Zhou H, Aebersold R: **Quantitative profiling of differentiation-induced microsomal proteins using isotope-coded affinity tags and mass spectrometry.** *Nat Biotechnol* 2001, **19:**946-951.
15. Martin SE, Shabanowitz J, Hunt DF, Marto JA: **Subfemtomole MS and MS/MS peptide sequence analysis using nano-HPLC micro-ESI fourier transform ion cyclotron resonance mass spectrometry.** *Anal Chem* 2000, **72:**4266-4274.